Original Paper

Responsiveness, Reliability, and Minimally Important and Minimal Detectable Changes of 3 Electronic Patient-Reported Outcome Measures for Low Back Pain: Validation Study

Robert Froud[1,2], PhD; Carol Fawkes[3], PhD; Jonathan Foss[1,4], PhD; Martin Underwood[1], MBChB, MD (UK); Dawn Carnes[3,5], PhD

[1]Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, United Kingdom

[2]Institute of Health Sciences, Kristiania University College, Oslo, Norway

[3]Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom

[4]Department of Computer Science, University of Warwick, Coventry, United Kingdom

[5]Faculty of Health, University of Applied Sciences and the Arts, Western Switzerland, Switzerland

Corresponding Author:
Robert Froud, PhD
Clinical Trials Unit
Warwick Medical School
University of Warwick
Gibbet Hill Campus
Coventry, CV4 7AL
United Kingdom
Phone: 44 024 7657 4221
Email: r.froud@warwick.ac.uk

Abstract

Background: The Roland Morris Disability Questionnaire (RMDQ), visual analog scale (VAS) of pain intensity, and numerical rating scale (NRS) are among the most commonly used outcome measures in trials of interventions for low back pain. Their use in paper form is well established. Few data are available on the metric properties of electronic counterparts.

Objective: The goal of our research was to establish responsiveness, minimally important change (MIC) thresholds, reliability, and minimal detectable change at a 95% level (MDC_{95}) for electronic versions of the RMDQ, VAS, and NRS as delivered via iOS and Android apps and Web browser.

Methods: We recruited adults with low back pain who visited osteopaths. We invited participants to complete the eRMDQ, eVAS, and eNRS at baseline, 1 week, and 6 weeks along with a health transition question at 1 and 6 weeks. Data from participants reporting recovery were used in MIC and responsiveness analyses using receiver operator characteristic (ROC) curves and areas under the ROC curves (AUCs). Data from participants reporting stability were used for analyses of reliability (intraclass correlation coefficient [ICC] agreement) and MDC_{95}.

Results: We included 442 participants. At 1 and 6 weeks, ROC AUCs were 0.69 (95% CI 0.59 to 0.80) and 0.67 (95% CI 0.46 to 0.87) for the eRMDQ, 0.69 (95% CI 0.58 to 0.80) and 0.74 (95% CI 0.53 to 0.95) for the eVAS, and 0.73 (95% CI 0.66 to 0.80) and 0.81 (95% CI 0.69 to 0.92) for the eNRS, respectively. Associated MIC thresholds were estimated as 1 (0 to 2) and 2 (–1 to 5), 13 (9 to 17) and 7 (–12 to 26), and 2 (1 to 3) and 1 (0 to 2) points, respectively. Over a 1-week period in participants categorized as “stable” and “about the same” using the transition question, ICCs were 0.87 (95% CI 0.66 to 0.95) and 0.84 (95% CI 0.73 to 0.91) for the eRMDQ with MDC_{95} of 4 and 5, 0.31 (95% CI –0.25 to 0.71) and 0.61 (95% CI 0.36 to 0.77) for the eVAS with MDC_{95} of 39 and 34, and 0.52 (95% CI 0.14 to 0.77) to 0.67 (95% CI 0.51 to 0.78) with MDC_{95} of 4 and 3 for the eNRS.

Conclusions: The eRMDQ was reliable with borderline adequate responsiveness. The eNRS was responsive with borderline reliability. While the eVAS had adequate responsiveness, it did not have an attractive reliability profile. Thus, the eNRS might be preferred over the eVAS for measuring pain intensity. The observed electronic outcome measures’ metric properties are within the ranges of values reported in the literature for their paper counterparts and are adequate for measuring changes in a low back pain population.

(J Med Internet Res 2018;20(10):e272) doi: 10.2196/jmir.9828

KEYWORDS

electronic patient-reported outcome measures; validation; responsiveness; reliability; minimally important change; minimal detectable change; Roland Morris Disability Questionnaire; visual analog scale; numerical rating scale

Introduction

Low back pain is a common and costly problem resulting in substantial personal, social, and economic burdens and is the number one cause of disability globally [1,2]. Low back pain is a symptom rather than a disease and most low back pain is nonspecific (ie, where no specific underlying cause has been identified, but where the term lacks formal definition and where definitions in trials have been diverse) [1,3]. The lifetime prevalence of low back pain is between 60% and 84% [4,5]. The global problem of low back pain is getting worse due to aging and increasing population size [6,7]. The number of clinical trials of interventions for low back pain has been increasing, with over 30 trials of interventions for low back pain now being published annually [8]. Patient-reported outcome measures (PROMs) in the form of paper questionnaires are typically used in these trials to judge the effectiveness of the health technology under investigation [8].

Disability and pain are by far the most commonly measured domains in trials of interventions for low back pain; each is measured at least twice as often as any other domain [8]. The visual analog scale (VAS) and numerical rating scale (NRS) are most commonly used for measuring pain intensity and the Roland Morris Disability Questionnaire (RMDQ) is most commonly used for measuring functional disability [8]. These are quasi-continuous measures where the relationship between the observed item responses and the unobserved latent variable is assumed to be consistent with a reflective conceptual framework [9]. There is evidence that paper forms of VAS and NRS have been in use since at least the early to mid-20th century, and the RMDQ has been used since 1983 [10-12].

The validity of a PROM is defined as "the degree to which an instrument truly measures the construct(s) it purports to measure" [13]. Several aspects that compose what we consider to constitute good development and validation of PROMs postdate the introduction of these particular instruments. Validation exercises have been performed retrospectively, results have accrued over time, and endorsement and use of the measures have survived the process [14-16]. Notwithstanding healthy academic debate, it is generally accepted that these outcome measures have reasonable face validity and content validity, and they have at times been considered the legacy gold standard for comparison for assessing the criterion/convergent validity of other instruments [17-19].

Measuring patient/participant change in health status using browser-based technology and mobile device technologies is a natural progression. Digital PROMs and ports of existing paper PROMs to digital media have become known as electronic patient-reported outcomes measures [20]. When migrating existing paper PROMs to electronic patient-reported outcome measures (ePROs), there are aspects relating to the metric validity of the instrument that may need to be reassessed. Some aspects of validity are clearly independent of whether the instrument is completed on paper or digitally—for example, the content wording (unless it is culturally or clinically out of date) and the extent to which this content is judged to appropriately span the domains of the health construct being measured (ie, content and face validity). However, other aspects of validity that relate directly to measurement performance should not be assumed to be unchanged.

For any instrument designed to measure change in a health construct, 2 properties are particularly relevant: reproducibility (ie, reliability) and responsiveness. Reliability is the extent to which the same results are obtained on repeated measures when no real change in health status has occurred [21,22]. An analogy using a bathroom scale is that it is desirable that the scale shows the same weight upon time-standardized daily measurement when there truly is no true change in a person's weight; if this is the case, the scale may be said to be reliable. Conversely, responsiveness is analogous to the scale detecting an important change when one truly exists. As users' physical interactions with ePRO versions of PROMs differs in fundamental respects from paper versions, we suggest that reassessing these 2 key change measurement properties is necessary before advocating their widespread use in health research.

In analyses of trials or evaluations of health interventions, using PROMs to decide when an individual participant has responded facilitates interpretation of intervention effect [23]. Responder analysis permits the number of improvements to simply be counted and compared by arm using several clear statistics. These are intuitive reporting methods, and there is consensus that back pain trials should incorporate these [23-25]. However, to be able to do this, it is necessary to know (1) the minimum thresholds considered important to an individual participant—the minimally important change (MIC)—and (2) what magnitudes of change can be detected beyond the inherent measurement error of the instrument—the minimal detectable change (MDC) [26,27]. These thresholds may be altered by the change in media from paper to digital and may also be population specific [28,29].

We aimed to determine reliability and responsiveness, MIC and MDC, for electronic versions of the VAS, RMDQ, and NRS as administered via Web browser and Android or iOS app to adults with low back pain who visit osteopaths.

Methods

Recruitment

We recruited adults with low back pain from osteopathic clinics in England and Wales. Participants were recruited by osteopaths on our behalf and provided with an enrollment code and instructions for installing the iOS or Android app (from the App

Store or Google Play) or completing the outcome measures using a Web browser.

We assumed an attrition rate of up to 70% and a recovery rate (ie, participants who indicate that they are much better or completely recovered using a health transition question) of over 90% in those with acute and subacute low back pain (ie, low back pain present for less than 3 months) [30]. Thus, for our responsiveness study, for which we required improved participants, we sought to recruit a minimum of 200 people with acute and subacute low back pain to ensure at least 50 eligible 6-week measurements. For people with chronic low back pain receiving manual therapy, we assumed up to the same rate of attrition but a lower rate of recovery (45%) [24]. For our test-retest study, we required stable participants who identified as remaining stable over a period of 1 week; thus, we sought to recruit 400 chronic patients to find 50 participants self-identifying as stable (ie, reporting no change on a health transition question). Participants were invited to complete the electronic versions of outcome measures at baseline, 1-week, and 6-week follow-up time points. Participants were offered a £5 (US $7) retail gift voucher for completing the outcome measures.

Software

We used Android and iOS apps and a Web app with an associated form builder that was developed by Clinvivo Ltd, a University of Warwick spin-out company [31]. The apps, which function identically across platforms, permitted PROMs to be typeset and then administered to patients securely on their own devices. Data in transit are encrypted using a Secure Sockets Layer, and data at rest are encrypted using a Rivest-Shamir-Aldeman and Advanced Encryption Standard encryption hybrid. At the end of the study period, data were encrypted using the open Pretty Good Privacy standard and transferred from Clinvivo to researchers. The iOS, Android, and Web apps sent data one way and did not receive or redisplay personal data. The platform presented an electronic version of the instrument and reminded participants to complete outstanding follow-up measurements, as appropriate. Off-line completion in apps was permitted in cases of interrupted connectivity, with submissions occurring upon restoration of connectivity. Reminders, which were received up to twice per follow-up measurement due, were sent directly to devices for app-enrolled participants and by email to Web-enrolled participants (up to 2 reminders).

Electronic Versions of Patient-Reported Outcome Measures

The VAS is a continuous scale running from 0 to 100 mm measuring current pain intensity [32]. It is the most commonly used outcome measure in trials of interventions for nonspecific low back pain overall [8]. Huskisson is commonly credited with its development in 1974; however, there is evidence that it was being used at least as far back as 1921 [11]. Intellectual property rights are in the public domain, and no permissions are required for use, reproductions, or modifications. Completion of the paper scale involves a person marking a line on the scale indicating their level of pain between 2 anchored scales that typically have wordings of “no pain” on the left (ie, 0 mm) and “worst possible pain” or “worst imaginable pain” on the right (ie, 100 mm) [33,34]. On paper, the distance of the marked line is then measured from the point of 0 pain and reported in mm. In migrating this to an electronic version (eVAS), we implemented a slider that could be dragged into position. We did not force the scale to render at 10 cm to allow for resizing to screens of different devices. Thus, we report scores in units rather than mm, where 1 unit is 1/100th of the scale (ie, where the pointer can be set at any one of 101 different positions) as rendered (Figure 1).

The RMDQ is a 24-item questionnaire measuring functional disability due to back pain that was developed in the early 1980s [10]. It is the most commonly used outcome measure in trials of interventions for low back pain overall [8]. The original paper version of the instrument is well established [35-38]. No permissions are required for its use, reproductions, or modifications [39]. Scores on the RMDQ range from 0 to 24, where higher scores indicate greater disability. Participants are given a statement with which they may indicate agreement by ticking a box. Participants are asked to tick statements that they feel describe them on that day and to leave blank boxes next to statements that they feel do not. The score is then the sum total of checked items. Our electronic (eRMDQ) migration is an exact copy using multiselect check boxes (Figure 2). One year into the research, we added a box stating “none of the above symptoms” for participants to confirm that none of the statements applied to them and to confirm 0 scores were genuine and not reflective of a skipped question.

The NRS is an 11-point ordinal scale measuring current pain intensity [40,41]. Validation of the paper version is well established [41-43]. It is the fourth most commonly used outcome in trials of interventions for low back pain overall [8]. It is well established, with intellectual property rights in the public domain. Scores on the NRS range from 0, which typically is anchored “no pain” and 10, which typically is anchored “worst pain possible.” Our electronic (eNRS) migration is an exact copy with these anchor wordings (Figure 3). As the range of responses is exhaustive, completion of the scale was required for submission.

Figure 1. Electronic visual analog scale for pain intensity showing 63 units of pain intensity.

Figure 2. Electronic Roland Morris Disability Scale showing a part score of 3 units.

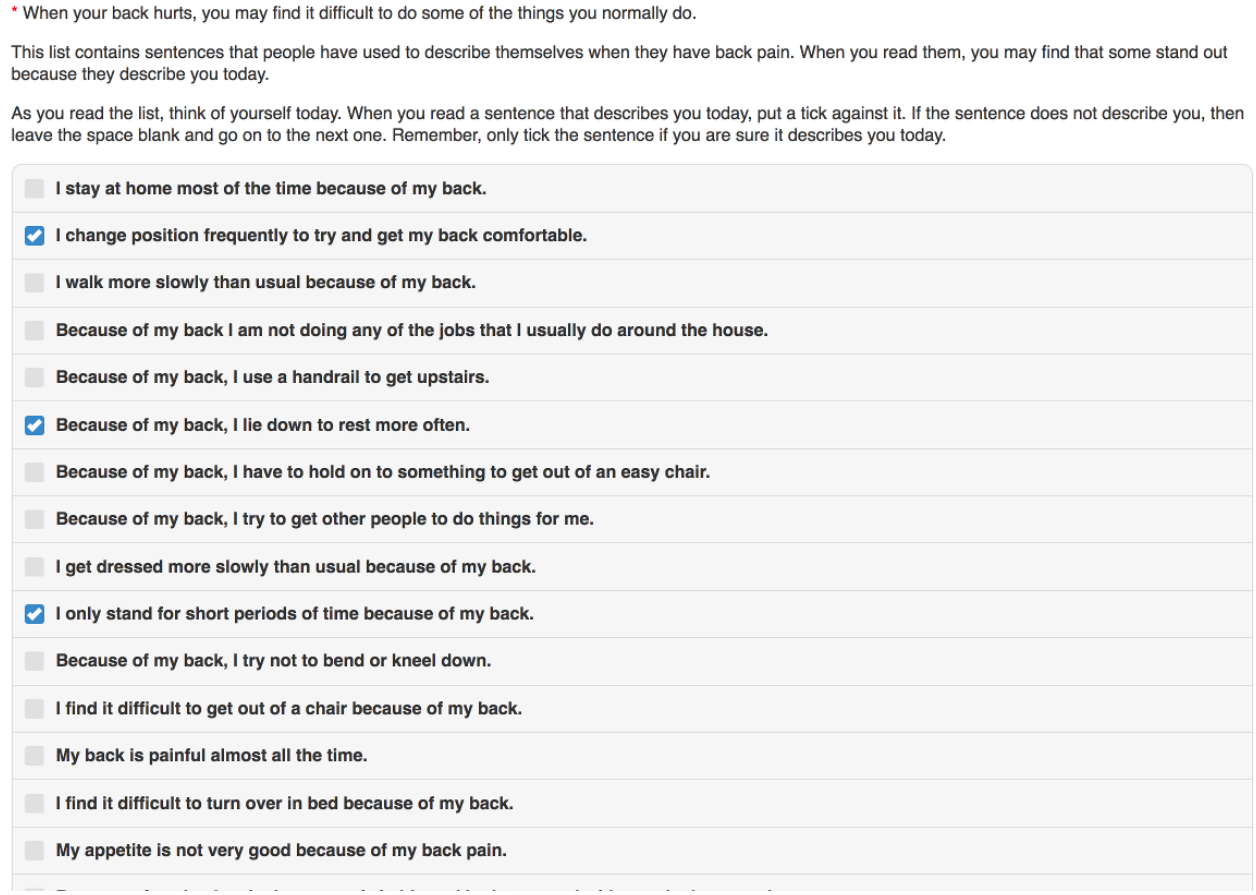
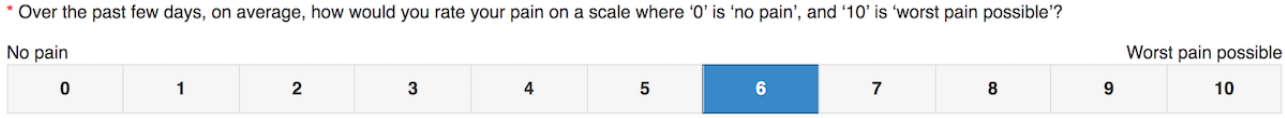


Figure 3. Electronic numerical rating scale for pain intensity showing a part score of 6 units.



Participants were also asked to electronically complete a health transition question at 1- and 6-week follow-up time points. The transition question was a single question with the wording "Overall, how would you rate the change in your symptoms since beginning this study?" where the participant could respond on a 7-point scale [44]: 1–completely recovered, 2–much improved, 3–slightly improved, 4–no change, 5–slightly worsened, 6–much worsened, and 7–vastly worsened.

Assessment

We aimed to have 50 completed paired measurements in improving participants for responsiveness assessments and 50 completed test-retest measurements in stable participants. We defined improving participants a priori as participants who selected much improved or completely recovered using the transition question. Improving participant scores were used to assess responsiveness at 1 and 6 weeks. For our test-retest study, we defined stable participants a priori as those who select no change at 1 week, and in the case of having too few observations, a post hoc sensitivity analysis including those who selected either slightly worsened, no change, or slightly improved. This alternative "about the same" approach to marking stability has been used elsewhere [45]. Allowing 1 week is typical in low back pain test-retest studies; clinically, this is close enough for the people with chronic pain to remain stable but far enough apart that participants cannot easily recall their initial responses. It was anticipated that the chronic population would predominantly contribute participants to the test-retest study and improving participants would come from across all chronicity subpopulations.

Statistical Analyses

To measure responsiveness in a way that is consistent with the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) definition, we constructed receiver operator characteristic (ROC) curves for 1- and 6-week data using a dichotomized transition question as the external criterion [22]. The area under the ROC curve (AUC) is then a metric of responsiveness, accepting that the external criterion reasonably includes the construct of interest [46]. The approach has previously been used to quantify responsiveness across all 3 paper versions of instruments [47]. ROC AUCs of over 0.70 were considered to be adequate [9,48]. We dichotomized the transition question such that participants responding completely recovered and much improved were

considered improved and all other responses were considered not improved.

We also used ROC curves and the transition question external criterion for 1- and 6-week data to quantify the MIC, which is defined as "the smallest [change] in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (see note 1 in Multimedia Appendix 1) [43,49]. We used a MIC estimator based on the minimum sums of squares method, which consistently selects the cut-point closest to the top left corner of ROC space, as required when sensitivity and specificity are valued equally [50]. We calculated confidence intervals for MIC point estimates using bootstrapping [51].

To estimate reliability, we calculated intraclass correlation coefficients (ICCs) [52,53]. ICC values usually range from 0 to 1 [54]. ICC values above 0.75 may be interpreted as excellent agreement, values of 0.40 to 0.75 indicate poor to fair agreement, and values of below 0.40 indicate poor agreement [55]. We calculated the standard error of measurement [53]. We used this to estimate the minimal detectable change at the 95% level (MDC_{95}) (see notes 2 to 4 in Multimedia Appendix 1) [53,56,57].

Transition questions can be highly correlated with follow-up score rather than change [24,43,58]. Guyatt et al [58] assert that if a transition question is truly measuring change then a correlation between the baseline score and transition question and the follow-up score and transition question should ideally be present, equal, and opposite. In addition, they suggest that in a linear regression model with follow-up score entered as the initial explanatory variable, the baseline score should explain a significant proportion of the residual variance in the transition rating [58]. We performed Pearson correlations and fitted regression models to explore the degree to which the transition question measured change or simply reflected follow-up status. Log rank tests were used to assess significance of the addition of baseline score.

All analyses were performed using Stata version 14.2 (StataCorp LLC). The program rocmic was used to estimate MIC and the ROC AUC, which for ROC AUC uses the lroc program [51,59].

Power and Sample Size

With the notable exception of construct validity, sample sizes in validation studies generally are not calculated based on power to test hypotheses: the estimation of reliability and responsiveness parameters is focused on the extent to which the coefficients describing these parameters approach 1 (which would represent perfect reliability/responsiveness) rather than their difference from 0 or some other null value. Generally, a sample size of at least 50 participants is considered adequate for this purpose [9,60]. Assuming an ICC of 0.7, with 50 participants we would be able to estimate the ICC to within a 95% CI of +/–0.14. Alternatively, for an ICC of 0.8, we would be able to estimate to within a 95% CI of +/–0.10 [9]. For responsiveness, with 50 participants and assuming an AUC of 0.8 and equal numbers of cases and noncases, we would be able to estimate AUC to within a 95% CI of +/–0.12 [61].

As standard errors (SEs) for MIC estimates are not readily calculable, we used bootstrapping to generate SEs and 95% CIs [51,62]. Previous simulation work on the paper-based RMDQ in a similar population suggested that 2500 bootstrap samples was sufficient to ensure SE convergence [63]. To explore whether this is the case for the eRMDQ (and also whether it is an appropriate number of replications for the eNRS and eVAS), we simulated SEs by randomly sampling n observations (with replacement) from our dataset for an increasing number of n, where n is an integer, beginning at 20 and increasing by increments of 20, up to 6000 [62,64]. We then graphically assessed SE convergence and used the point of convergence to inform the number of bootstrap replications.

Data Exclusions, Assumptions, and Variations

Prior to the addition of the "none" box, we imputed 0 scores for all baseline submissions with no eRMDQ boxes ticked and assumed and imputed a 0 score for eRMDQ follow-up scores in the case that the baseline eRMDQ score was greater than 0 and a submission had been made for the follow-up period in question. When the eVAS rendered, it did so with the slider in the 0 position. In the case of a submission for an untouched eVAS, a score of 0 was assumed valid. The eNRS was a required response and necessitated a selection for submission.

As part of the basic demographic details collected, we included a list of presenting complaints, featuring low back pain among 15 other common musculoskeletal presentations and the opportunity to report a complaint not listed in a free-text box. The list of complaints was derived from earlier survey work developed as part of a national data collection initiative [65,66]. We excluded all cases where a participant had not checked the low back pain box (data from non-low back pain cases were used in unrelated research).

Ethics Approval

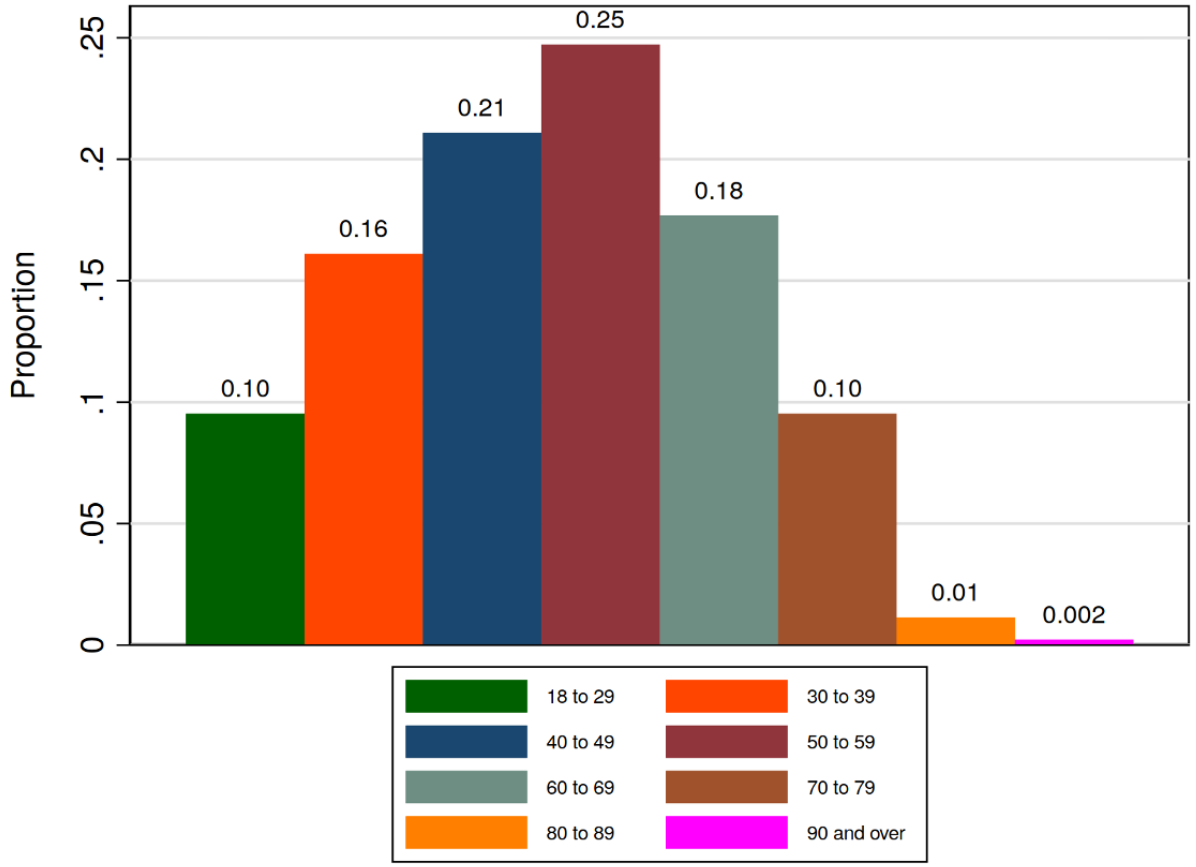
Ethics approval was obtained from the research ethics committee at Queen Mary University of London (QMERC2014/18).

Results

User Statistics and Demographics

We collected data from 575 people from 30 osteopathic clinics between July 15, 2014, and May 3, 2017. Of these, 442 (76.9%) reported low back pain as their main complaint. The average submission time for 1-week scores was 7.4 (SD 0.79) days after baseline. The average submission time for 6-week scores was 42.5 (SD 0.9) days after baseline. Of the participants, 60.4% (267/442) were female, 69.2% (306/442) identified as being in full or part-time employment, 1.1% (5/442) were long-term sick, 3.6% (16/442) identified as looking after home/family, 19.7% (87/442) were retired, 1.4% (6/442) were in full-time education, 2.9% (13/442) were unemployed, and 2.0% (9/442) selected other or preferred not to disclose. Figure 4 shows a histogram of patient-reported age at baseline.

Figure 4. Histogram of patient age at baseline.



We collected baseline eNRS data from 442 participants, and we collected baseline eVAS and eRMDQ data from 247 participants. One-week data were collected from 187 and 97 participants, respectively, and 6-week data were collected from 91 and 40 participants, respectively. Figure 5 shows the incidence of recovery in these groups. There was 1 missing data point for eNRS at baseline (0.2%) and 1 week (0.5%) for which we were unable to confirm cause. Table 1 summarizes ePRO submission scores using median and interquartile range and Table 2 summarizes recoveries and cumulative recoveries recorded using the transition question. Change scores (not shown) more closely followed normal distributions.

The addition of baseline score generally explained a significant proportion of the variance in the transition question over and above follow-up score. The transition question correlated with follow-up score but not with baseline score. Comprehensive results for the Guyatt analyses on the transition question's performance in measuring change are listed in note 2 in Multimedia Appendix 1.

Evaluation Outcomes

Graphically, SE convergence appeared to be asymptotically complete at around 5000 bootstrap replications (Figure 6); thus 5000 replications were used to generate confidence intervals for the MIC estimates in Table 3. Responsiveness point estimates (Table 3) were borderline adequate (AUC≈0.7) or above adequate for all instruments and time points. The AUC confidence interval for the RMDQ at 6 weeks spanned the null value (Table 3).

Using no change as a criterion for judging stability, we did not achieve our a priori threshold of 50 test-retest data points for comparison across any of the instruments. Of the people who said they had no change at 1 week, 65% (15/23) had chronic pain. Allowing slightly improved and slightly worsened to count as stable enabled us to achieve this threshold for the eNRS only. Of people who said they had no or slight change at 1 week, 63% (53/84) had chronic pain. Notwithstanding the lack of data, the eRMDQ reliability (agreement) was excellent using either analysis, with CIs spanning fair to excellent in both analyses (Table 4). For the eVAS per protocol analysis, the agreement was fair with CIs spanning poor to fair, and in the sensitivity analysis, the agreement was poor to fair with a CI range spanning poor to fair (Table 4). For the eNRS per protocol analysis, the agreement was poor to fair with a CI spanning poor to excellent, and for the sensitivity analysis, agreement was fair with a CI spanning poor to fair to excellent (Table 4).

Figure 5. Flowchart showing completion rates at 1 and 6 weeks, chronicity status, and the incidence of self-reported recovery using the health transition question for participants who also completed the electronic numerical rating scale, and electronic Roland Morris Disability Questionnaire, and electronic visual analog scale measurement.

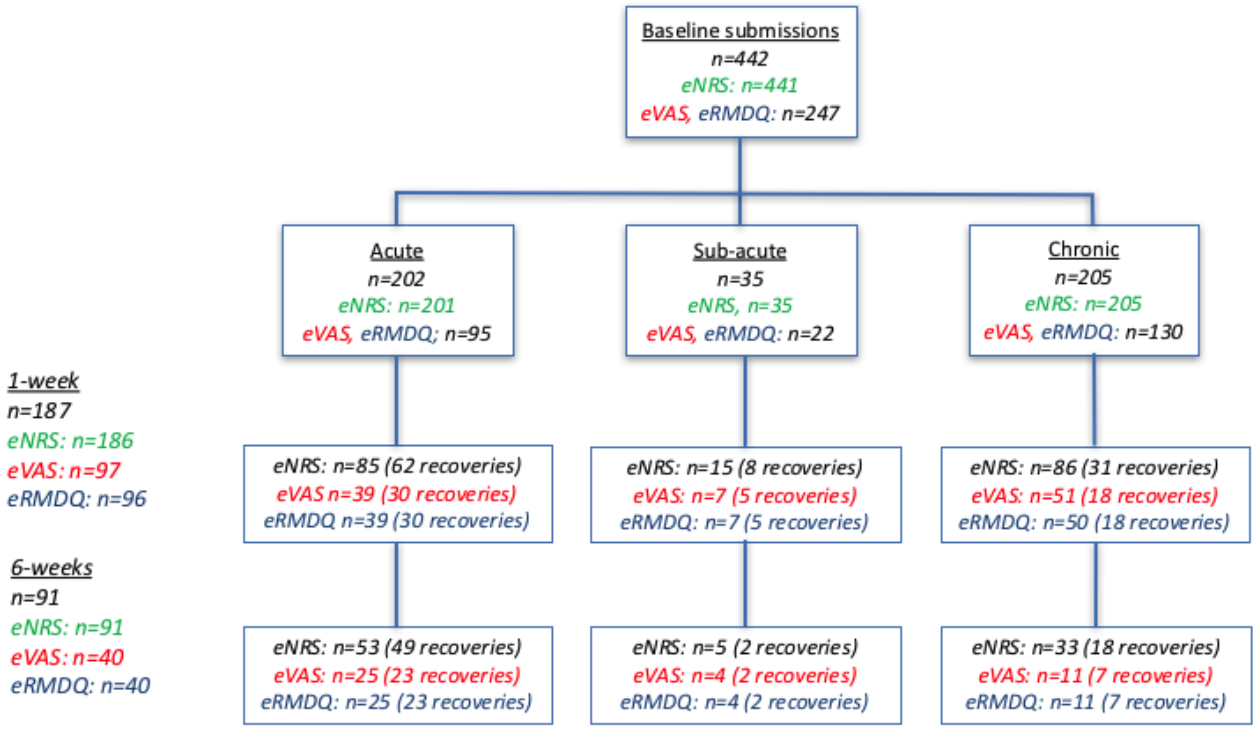


Table 1. Baseline, 1-week, and 6-week scores across the whole sample.

Score	Baseline		1 week		6 week	
	Median (IQR)	n[a]	Median (IQR)	n	Median (IQR)	n
eRMDQ[b]	4 (6)	247	2 (4)	96	2 (3.5)	40
eVAS[c]	41 (32)	247	24 (19)	97	19 (19)	40
eNRS[d]	5 (4)	441	3 (3)	186	2 (2)	91

[a]The number of received measurements at 1 week and at 6 weeks, respectively.

[b]eRMDQ: electronic Roland Morris Disability Questionnaire.

[c]eVAS: electronic visual analog scale.

[d]eNRS: electronic numerical rating scale.

Table 2. Recoveries and cumulative recoveries recorded using the transition question

Transition question	n	Recoveries, n (%)	Cumulative recoveries, n[a] (%)
1 week	187	101 (54)	101 (23)
6 weeks	91	69 (76)	170 (38)

[a]Where the frequency of cumulative recoveries are shown as a proportion of all 442 baseline participants.

Figure 6. Graphs showing minimally important change bootstrap standard error convergence from simulations with increasing replication numbers. MIC: minimally important change, NRS: numerical rating scale, RMDQ: Roland Morris Disability Questionnaire, VAS: visual analog scale.

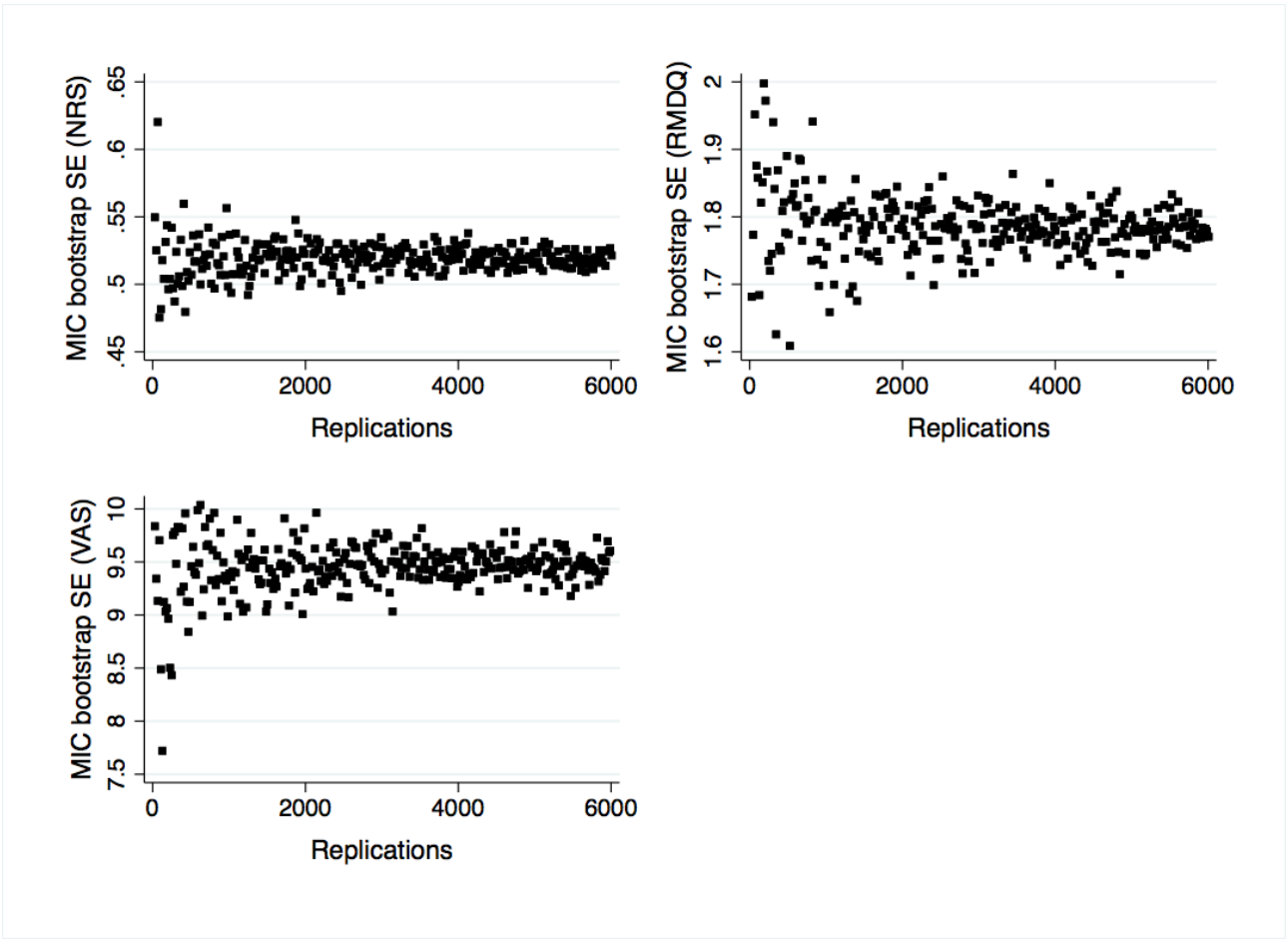


Table 3. Responsiveness and minimally important change by instrument and 1-week and 6-week follow-up time periods.

Instrument and time period	Receiver operator characteristic AUC[a]	95% CI	n[b]	Minimally important change points/ eVAS[c] units (% of baseline score)	95% CI
eRMDQ[d]					
1 week	0.69	0.59 to 0.80	96	1 (19)	0 to 2
6 weeks	0.67	0.46 to 0.87	40	2 (38)	–1 to 5
eVAS					
1 week	0.69	0.58 to 0.80	93	13 (32)	9 to 17
6 weeks	0.74	0.53 to 0.95	40	7 (17)	–12 to 26
eNRS[e]					
1 week	0.73	0.66 to 0.80	185	2 (43)	1 to 3
6 weeks	0.81	0.69 to 0.92	91	1 (21)	0 to 2

[a]AUC: area under the curve.

[b]The number of change scores available (ie, from available pairs of measurements at baseline and follow-up time point) at 1 week and 6 weeks, respectively.

[c]eVAS: electronic visual analog scale.

[d]eRMDQ: electronic Roland Morris Disability Questionnaire.

[e]eNRS: electronic numerical rating scale.

Table 4. Intraclass correlation coefficients from test-retest study in a per protocol stable sample and a pseudo-stable sample with associated minimal detectable change thresholds.

Instrument and condition	n[a]	Intraclass correlation coefficient$_{agreement}$	95% CI	MDC$_{95}$[b] points/eVAS[c] units
eRMDQ[d]				
Per protocol	15	0.87	0.66 to 0.95	4
Allowing slight change	43	0.84	0.73 to 0.91	5
eVAS				
Per protocol	15	0.31	–0.25 to 0.71	39
Allowing slight change	43	0.61	0.36 to 0.77	34
eNRS[d]				
Per protocol	22	0.52	0.14 to 0.77	4
Allowing slight change	83	0.67	0.51 to 0.78	3

[a]The number of cases satisfying the condition for analysis as a stable case.

[b]MDC$_{95}$: minimal detectable change at the 95% level.

[c]eVAS: electronic visual analog scale.

[d]eRMDQ: electronic Roland Morris Disability Questionnaire.

[e]eNRS: electronic numerical rating scale.

Discussion

Principal Findings

The results suggest that the eRMDQ had borderline adequate responsiveness and excellent reliability. Conversely, the eNRS had relatively good responsiveness at 6 weeks but borderline adequate reliability. The eNRS outperformed the eVAS, which had adequate responsiveness but relatively poor reliability. As test-retest numbers were few, eVAS CIs spanned poor to excellent, and thus further investigation is warranted. While exploring use by age was not a specific study objective, we note the results indicate encouraging use by older people from this population.

Comparison With Prior Work

Across acute and chronic back pain populations there has been like-for-like evaluation (ie, using similar and directly comparable methods) of the properties of paper versions of the outcome measures explored. ROC AUC for the RMDQ ranges from 0.64 to 0.93 [45,47,67-75]. ROC AUC for the NRS ranges from 0.67 to 0.93 [41,42,47,67,75,76]. ROC AUC for the VAS ranges from 0.71 to 0.93 [47,72,77-79]. Our results are within these ranges at 6 weeks for all instruments and for all but our eVAS instrument at 1 week, where our point estimate approaches the lower border of the range. Our eVAS data are nevertheless consistent with the range (ie, insofar as the upper CI overlaps). Estimates of ROC AUC for the VAS are fewer in the literature, which might explain why the range of reported results is narrower than it is for the RMDQ and NRS.

MIC thresholds for RMDQ ranged between 1.5 and 5.0 [21,24,35,67,68,72,75,80-83], for the NRS between 1.5 and 4.0 [41-43,67,75,81,84], and for the VAS between 15 and 28 mm [72]. Our absolute MIC thresholds are comparable but are toward the lower side of this range. MIC estimates are known to increase with baseline severity, and relatively low baseline scores likely explain our relatively low thresholds [68,75,81,84]. However, MIC thresholds in our results, expressed as percentage change from baseline, average 28% across all 3 instruments and all time points. This is consistent with the suggestion of Ostelo et al [29] (following their review of MIC and MDC literature) for using an improvement of between 20% and 30% of baseline score for the RMDQ, NRS, and VAS as a MIC threshold. We emphasize that the MIC thresholds relate to the degree of change that may be considered important for an individual and not what degree of difference may be considered important at a population level [27,85,86]. We note that the 2 negative CIs imply consistency of the data, with the true MIC thresholds being in the opposite direction of improvement (ie, a slight deterioration). This is likely an artifact of low power, and we suggest using inflated sample sizes for future studies based on the bootstrapped standard error observations.

Reported ICC estimates for the RMDQ have ranged from 0.42 to 0.95 [45,67,81,87] and for the NRS from 0.92 to 0.98 [67,81], and an estimate for the VAS of 0.71 has been reported [88]. Our results are within the ranges reported, but our ICC point estimate for the eVAS is lower than the reported paper VAS estimate. It is conceivable that rendering the eVAS slider in a 0 position might lead to additional variance in the case that the outcome is overlooked (ie, leading to a comparatively lower ICC), and future research might explore whether a touch to confirm 0 design is acceptable to users. We also note that some of the ICC values in the literature ranges may have been derived from ICCs for consistency rather than agreement; this is a practice known to exist (although it is not always clear which approach has been used) and known to overestimate reliability [53].

MDC$_{95}$ estimates reported (or in the case of the NRS only, either reported or calculated from reported standard error of measurements) have ranged from 5.0 to 12.1 for the RMDQ [21,24,35,45,56,67,81,83], from 2.4 to 11 (ie, almost the full

width of the scale) for the NRS [41,45,67,81,84], and from 21.0 to 33.5 for the VAS [79,88,89]. Our estimates are slightly better than average for the RMDQ, toward the lower end of the range for the NRS, and comparable to the available estimates for the VAS.

In terms of comparison to studies assessing these instruments as ePROs, Bird et al [90] conducted a test-retest study among 22 healthy adults of the VAS administered on a tablet and found ICCs of 0.90 (0.82 to 0.95) as compared to 0.96 (0.92 to 0.98) in a paper version that participants completed simultaneously. It is difficult to compare the results with this study, as the time between test and retest was less than 30 minutes. A much shorter period between test and retest might be appropriate in some populations (eg, where change in acute pain must be measured over short spaces of time). In these cases, participants may be more prone to panel conditioning, where the second response is affected by recall of the first response [91]. For back pain, most interventions focus on chronic pain and longer time periods. When exploring reliability of low back pain outcome measures, a 1-week gap between test and retest is typical. Bijur et al [92] and Gallagher et al [93] have used small time frames between tests on a paper-based VAS in acute pain populations and demonstrate similarly high ICCs of 0.97 (0.96 to 0.98) and 0.99 (0.989 to 0.992), respectively. Also of relevance but not directly comparable is work by Bishop et al [94], who administered the RMDQ on paper and online and constructed limits of agreement, demonstrating equivalence with a score difference of only 0.03 points and a Bland-Altman range of –2.77 to 2.83.

Finally, we note that the distribution of the user age of the health outcomes app in this population appears to be higher than the age of health app users [95].

Implications

None of our results differs materially from ranges observed in population-similar and methodologically alike studies of paper counterparts. There is thus some suggestion that the ePROs under evaluation are suitable substitutes for PROMs for measuring change in low back pain. The eNRS outperformed the eVAS in terms of responsiveness and reliability. As such, we suggest the eNRS might be preferred over the eVAS for the measurement of low back pain intensity, but we caution that subsequent confirmatory research is warranted.

Limitations

The principal limitation is that in several cases we had small sample sizes. We had intended to recruit sufficient numbers to have at least 50 people for each assessment, in line with recommendations, but we failed to meet these targets, mainly as we underestimated the incidence of stability, although we also underestimated attrition [9]. There were high rates of improvement in people receiving treatment, and this is a hazard of nesting a test-retest design within a protocol where participants are receiving routine clinical treatment. This was of consequence in the eRMDQ responsiveness analysis, where the data are consistent with a null population parameter and thus 6-week responsiveness of the eRMDQ requires confirmation in a larger sample. Having too few data has greater implications for the test-retest assessment of the VAS where the CIs span coefficient values that can be interpreted at their extremes as either poor or excellent. It is less of an issue for the eRMDQ because while the numbers are low and lower at 1 and 6 weeks, respectively, the stronger signal combined with boundary proximity leads to narrower and more useful CIs.

It is not ideal that we permitted slightly worse and slightly improved categories to indicate stability in our test-retest, although we note a similar approach has been observed previously [45]. Further, this was a post hoc decision taken in light of having too few observations to use our more stringent a priori criterion of including only those reporting no change. The results using our a priori approach but with few observations are offered as sensitivity analyses that may provide useful comparison.

Having relatively few observations also meant that we were unable to explore differences by platform (ie, iOS, Android, and Web browser) or explore MIC as a function of baseline score (eg, stratifying by number in category of severity) or separately by chronicity, which may have been useful and allowed us to explore any differences in these metrics by chronicity. Thus, our focus here is pragmatic and results are generalizable to the population of adults with low back pain who consult osteopaths, notwithstanding chronicity.

We recorded in our database only the summed eRMDQ score rather than individual responses. Had we retained detail of individual response profiles of the eRMDQ, we could have also calculated internal consistency (as well as aspects of modern test theory: Rasch analysis to examine item performance or factor analysis to explore data dimensionality). Whereas COSMIN conflate internal consistency with reliability in their taxonomy [22,96], we consider internal consistency to be an indication of the unidimensionality of a scale and of item redundancy rather than the degree to which a scale is free from measurement error. As such, and with respect to the reliability definition, we preferred to consider it separately. We had not immediately considered that the media used for completion might affect internal consistency or item functioning of a scale. On reflection, however, we think that it is conceivable that presenting the scale digitally may alter the way patients respond in such a way that these could be affected. Additionally, there may be self-selection effects of those more familiar with digital media joining the study, and this may be a factor that could be confounded with how a person responds.

It is not ideal that our transition question correlates with follow-up score but not with baseline score. This is emerging to be the case generally and is not something particular to evaluating electronic outcome measures [24,43,58]. This emergence in our view raises the more general question of whether it is appropriate to use transition questions at all to evaluate change in outcome measures. Apart from being overly driven by follow-up score, the assumption that the transition question is sufficiently driven by the same latent construct as the PROM, to the extent that it may be considered a gold standard, may be unrealistic. We have previously explored what people think about when they complete the transition question and what they think about when they complete the paper RMDQ

version, and we found discordance [97]. Pain appears to be a greater driver of the transition question, and the wording of the transition question (ie, attempting to place focus specifically on function or an explicit domain) does not appear to matter. In our study, we used the term symptoms. However, in the case that the suggestion arising from our previous research is incorrect, using a generic wording in the transition question might have the advantage of not favoring any one ePRO over another but the disadvantage of disassociating the transition question from any specific latent health construct. Use of a generically worded transition question would then introduce some information bias—for example, if people systematically attend more to a particular domain upon reading the word symptoms. We caution that the logic of the typically taken approach of using one outcome measure as a proxy gold standard of recovery and then using this proxy to judge domain-specific responsiveness and MIC thresholds in another may be questionable where there is domain mismatch.

There was a small amount of missing data at baseline and 1 week (a person in each case), which should have been impossible because a selection on the eNRS was a required response. We are uncertain of the cause but we suspect this might have been due to use of an obscure and/or obsolete browser.

This research was conducted solely in private care and people who pay to see osteopaths may differ from those attending publicly funded health care, as is more routinely the case in health services research. We note a lower than typical baseline severity (as compared to clinical trials) and thus some caution is indicated before generalizing to typical trial populations. Finally, our focus here was on the most commonly used domains and outcome measures in trials. The VAS is most commonly used overall (pain), RMDQ second most common (disability), and the NRS fourth most commonly used (pain). We did not include the third most commonly used outcome, the Oswestry Disability Questionnaire, which also measures disability [8]. Unlike the VAS and NRS, which are both single-item instruments, including two full disability questionnaires risked being unduly burdensome for participants. Qualitative work suggests that participants would prefer to spend only 5 to 10 minutes completing ePROs [98,99]. Including a direct comparison with paper versions would have permitted direct exploration of criterion validity; however, this approach would likely have been affected by panel condition and further added to participant burden.

Recommendations for Future Research

Sampling stable participants from people receiving routine clinical treatment allows the nesting of a test-retest design and makes for an efficient design. However, it produces some challenges for achieving sufficient recruitment over a realistic time period. It assumes that the transition question classification of unchanged is valid. As data suggest that transition question is driven more by follow-up state than change, the approach has some limitations. It would be scientifically preferable that test-retest studies are conducted within untreated populations. However, this has ethical and practical implications. When planning to nest a test-retest design within any treatment-containing protocol, based on rates observed in this study (using the lower eNRS no chance incidence), we recommend planning a study that is around 3 times larger (ie, seeking approximately 1200 people to obtain 50 stable participants). For study of responsiveness alone, about 250 participants should be sufficient to achieve 50 improvements at 6 weeks. The most extreme MIC threshold we estimated was 7 units (–12 to 26) for the eVAS at 6 weeks. This is lower than has been noted in studies of paper counterparts. Assuming the point estimate is representative of the population parameter, approximately 300 participants would be required to power a study to confirm the finding.

Retaining data at item level in future studies will permit more sophisticated analytics. There may need to be a cultural change as we transition from paper to digital measurement. The ability to more easily retain greater data resolution is a clear advantage of digital measurement and one that would be sensible to exploit. Further advantages in terms of cost, logistics, form validation, reminders, time logging, environmental factors, and reach are undeniable and, in our view, make electronic health measurement very attractive. More generally, routine outcome measurement in clinical practice may facilitate so-called learning health care systems and should be a shared goal of stakeholders across health care [100,101]. To achieve this, greater collaboration may be needed between clinicians, informatics specialists, and policy makers. We also encourage further metric testing of electronic versions of these and other legacy PROMs so that results may inform health services researchers and clinicians' choices of measure.

Conclusion

Each of the electronic outcome measures has metric properties that do not materially differ from values reported in the literature for their paper counterparts. A possible exception may be the reliability of the eVAS, for which there is insufficient existing research to make useful comparisons between paper and digital versions. The eRMDQ is adequate for measuring back-related disability, and the eNRS is adequate for measuring pain intensity. The eNRS should be preferred over the eVAS for the measurement of pain intensity.

Acknowledgments

Part of RF and JF's time on the study was funded by the Warwick Impact Fund, which administers the Higher Education Innovation Funding grant. The study was sponsored by University of Warwick but conducted from Queen Mary University of London, which sponsored the remainder of CF's doctoral research. Neither the sponsor nor funder had any involvement in study design, analysis, or reporting of results.

Authors' Contributions

RF and MU conceived of the study and applied for and were awarded the funding to do the study. CF undertook the day-to-day management and submitted the documents for consideration by the Queen Mary University of London ethics committee. JF and RF were responsible for data management. RF performed all analyses. RF wrote the first draft of the paper. All authors commented on and approved the manuscript.

Conflicts of Interest

RF, MU, and JF are directors and shareholders of Clinvivo Ltd, the University of Warwick spin-out company that provided the software for data collection in this study. The Higher Education Innovation Funding grant paid for the development of intellectual property licensed to Clinvivo and used in this study and also paid for UK retail vouchers used as incentives to recruit participants into the study. RF and DC are nonpracticing osteopaths; CF is a practicing osteopath. MU was chair of the National Institute for Health and Care Excellence accreditation advisory committee, for which he received a fee, until March 2017. MU is chief investigator or co-investigator on multiple previous and current research grants from the UK National Institute for Health Research (NIHR), Arthritis Research UK, Arthritis Australia, and the Australian National Health and Medical Research Council. He has received travel expenses for speaking at conferences from professional organizations hosting the conferences. He is an editor of the NIHR journal series for which he receives a fee. RF and MU have published multiple papers on chronic pain, some of which are referenced in this paper. RF, MU, and JF are part of an academic partnership with Serco Ltd related to return-to-work initiatives.

Multimedia Appendix 1

Technical notes and extended technical results for Guyatt analyses.

[PDF File (Adobe PDF File), 18KB-Multimedia Appendix 1]

References

1. Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, Lancet Low Back Pain Series Working Group. What low back pain is and why we need to pay attention. Lancet 2018 Mar 20:1. [doi: 10.1016/S0140-6736(18)30480-X] [Medline: 29573870]
2. Froud R, Patterson S, Eldridge S, Seale C, Pincus T, Rajendran D, et al. A systematic review and meta-synthesis of the impact of low back pain on people's lives. BMC Musculoskelet Disord 2014 Feb 21;15:50 [FREE Full text] [doi: 10.1186/1471-2474-15-50] [Medline: 24559519]
3. Amundsen PA, Evans DW, Rajendran D, Bright P, Bjørkli T, Eldridge S, et al. Inclusion and exclusion criteria used in non-specific low back pain trials: a review of randomised controlled trials published between 2006 and 2012. BMC Musculoskelet Disord 2018 Apr 12;19(1):113 [FREE Full text] [doi: 10.1186/s12891-018-2034-6] [Medline: 29650015]
4. Airaksinen O, Brox JI, Cedraschi C, Hildebrandt J, Klaber-Moffett J, Kovacs F, COST B13 Working Group on Guidelines for Chronic Low Back Pain. European guidelines for the management of chronic nonspecific low back pain. Eur Spine J 2006 Mar;15 Suppl 2:S192-S300 [FREE Full text] [doi: 10.1007/s00586-006-1072-1] [Medline: 16550448]
5. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet 2017 Feb 18;389(10070):736-747. [doi: 10.1016/S0140-6736(16)30970-9] [Medline: 27745712]
6. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017 Sep 16;390(10100):1211-1259 [FREE Full text] [doi: 10.1016/S0140-6736(17)32154-2] [Medline: 28919117]
7. Clark S, Horton R. Low back pain: a major global challenge. Lancet 2018 Mar 20:1. [doi: 10.1016/S0140-6736(18)30725-6] [Medline: 29573869]
8. Froud R, Patel S, Rajendran D, Bright P, Bjørkli T, Buchbinder R, et al. A systematic review of outcome measures use, analytical approaches, reporting methods, and publication volume by year in low back pain trials published between 1980 and 2012: respice, adspice, et prospice. PLoS One 2016;11(10):e0164573 [FREE Full text] [doi: 10.1371/journal.pone.0164573] [Medline: 27776141]
9. de Vet H, Terwee C, Mokkink L, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
10. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 1983 Mar;8(2):141-144. [Medline: 6222486]
11. Hayes M, Patterson D. Experimental development of the graphic rating method. Psychol Bull 1921;18:98-99.
12. Leavell H. Contributions of the social sciences to the solution of health problems. N Engl J Med 1952 Dec 04;247(23):885-897. [doi: 10.1056/NEJM195212042472305] [Medline: 13002642]
13. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010 Jul;63(7):737-745. [doi: 10.1016/j.jclinepi.2010.02.006] [Medline: 20494804]

14. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research: a proposal for standardized use. Spine (Phila Pa 1976) 1998 Sep 15;23(18):2003-2013. [Medline: 9779535]
15. Dworkin R, Turk D, Farrar J, Haythornthwaite J, Jensen M, Katz N, IMMPACT. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005 Jan;113(1-2):9-19. [doi: 10.1016/j.pain.2004.09.012] [Medline: 15621359]
16. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Report of the NIH task force on research standards for chronic low back pain. Spine J 2014 Aug 01;14(8):1375-1391. [doi: 10.1016/j.spinee.2014.05.002] [Medline: 24950669]
17. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, et al. The Quebec Back Pain Disability Scale. Measurement properties. Spine (Phila Pa 1976) 1995 Feb 01;20(3):341-352. [Medline: 7732471]
18. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain 2004 Mar;5(2):133-137. [doi: 10.1016/j.jpain.2003.12.005] [Medline: 15042521]
19. Nishiwaki M, Takayama M, Yajima H, Nasu M, Kong J, Takakura N. The Japanese version of the Massachusetts General Hospital acupuncture sensation scale: a validation study. Evid Based Complement Alternat Med 2017;2017:7093967 [FREE Full text] [doi: 10.1155/2017/7093967] [Medline: 28676831]
20. Wallwiener M, Matthies L, Simoes E, Keilmann L, Hartkopf AD, Sokolov AN, et al. Reliability of an e-PRO tool of EORTC QLQ-C30 for measurement of health-related quality of life in patients with breast cancer: prospective randomized trial. J Med Internet Res 2017 Sep 14;19(9):e322 [FREE Full text] [doi: 10.2196/jmir.8210] [Medline: 28912116]
21. de Vet HC. Reproducibility and responsiveness of evaluative outcome measures. Int J Technol Assess Healthc 2001;17(4):479-487. [Medline: 11758292]
22. Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, et al. COSMIN Checklist Manual version 9. 2012. URL:http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf [accessed 2018-08-29] [WebCite Cache ID 722fv3jFU]
23. Froud R, Underwood M, Carnes D, Eldridge S. Clinicians' perceptions of reporting methods for back pain trials: a qualitative study. Br J Gen Pract 2012 Mar;62(596):e151-e159 [FREE Full text] [doi: 10.3399/bjgp12X630034] [Medline: 22429424]
24. Froud R, Eldridge S, Lall R, Underwood M. Estimating the number needed to treat from continuous outcomes in randomised controlled trials: methodological challenges and worked example using data from the UK Back Pain Exercise and Manipulation (BEAM) trial. BMC Med Res Methodol 2009 Jun 11;9:35 [FREE Full text] [doi: 10.1186/1471-2288-9-35] [Medline: 19519911]
25. Froud R, Eldridge S, Kovacs F, Breen A, Bolton J, Dunn K, et al. Reporting outcomes of back pain trials: a modified Delphi study. Eur J Pain 2011 Nov;15(10):1068-1074. [doi: 10.1016/j.ejpain.2011.04.015] [Medline: 21596600]
26. de Vet HC, Terwee C, Ostelo R, Beckerman H, Knol D, Bouter L. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006;4(54):1. [Medline: 16925807]
27. Froud R, Underwood M, Eldridge S. Improving the reporting and interpretation of clinical trial outcomes. Br J Gen Pract 2012 Oct;62(603):e729-e731 [FREE Full text] [doi: 10.3399/bjgp12X657008] [Medline: 23265234]
28. Henschke N, van Enst A, Froud R, Ostelo RWG. Responder analyses in randomised controlled trials for chronic low back pain: an overview of currently used methods. Eur Spine J 2014 Apr;23(4):772-778 [FREE Full text] [doi: 10.1007/s00586-013-3155-0] [Medline: 24419902]
29. Ostelo R, Deyo R, Stratford P, Waddell G, Croft P, Von Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine 2008;33(1):90-94. [Medline: 18165753]
30. van Tulder M, Becker A, Bekkering T, Breen A, del Real MTG, Hutchinson A, COST B13 Working Group on Guidelines for the Management of Acute Low Back Pain in Primary Care. Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J 2006 Mar;15 Suppl 2:S169-S191 [FREE Full text] [doi: 10.1007/s00586-006-1071-2] [Medline: 16550447]
31. Clinvivo. 2018. URL:http://www.clinvivo.com/ [accessed 2018-01-12] [WebCite Cache ID 6wPydLMrk]
32. Huskisson E. Measurement of pain. Lancet 1974 Nov 09;2(7889):1127-1131. [Medline: 4139420]
33. Hägg O, Fritzell P, Nordwall A, Swedish Lumbar Spine Study Group. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 2003 Feb;12(1):12-20. [doi: 10.1007/s00586-002-0464-0] [Medline: 12592542]
34. Scott J. Vertical or horizontal visual analogue scales. Ann Rheum Dis 1979;38(560):1. [Medline: 317239]
35. Ostelo RWJG, de Vet HC, Knol DL, van den Brandt PA. 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. J Clin Epidemiol 2004 Mar;57(3):268-276. [doi: 10.1016/j.jclinepi.2003.09.005] [Medline: 15066687]
36. Dunn KM, Cherkin DC. The Roland-Morris Disability Questionnaire. Spine (Phila Pa 1976) 2007 Jan 15;32(2):287. [doi: 10.1097/01.brs.0000249551.00481.3d] [Medline: 17224833]
37. Roland M, Morris R. A study of the natural history of low-back pain. Part II: development of guidelines for trials of treatment in primary care. Spine (Phila Pa 1976) 1983 Mar;8(2):145-150. [Medline: 6222487]

38. Beurskens AJ, de Vet HC, Köke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996 Apr;65(1):71-76. [Medline: 8826492]
39. Roland-Morris Disability Questionnaire. 1983. URL:http://www.rmdq.org/ [accessed 2018-01-12] [WebCite Cache ID 6wPzCvX6M]
40. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. Ann Rheum Dis 1978 Aug;37(4):378-381 [FREE Full text] [Medline: 686873]
41. Childs J, Piva S, Fritz J. Responsiveness of the numeric pain rating scale in patients with low back pain. Spine 2005;30(11):1. [Medline: 15928561]
42. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001 Nov;94(2):149-158. [Medline: 11690728]
43. de Vet HC, Ostelo R, Terwee C, van der Roer N, Knol D, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 2007;16:131-142. [Medline: 17033901]
44. Lauridsen HH, Hartvigsen J, Korsholm L, Grunnet-Nilsson N, Manniche C. Choice of external criteria in back pain research: does it matter? Recommendations based on analysis of responsiveness. Pain 2007 Sep;131(1-2):112-120. [doi: 10.1016/j.pain.2006.12.023] [Medline: 17276006]
45. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002 Jan;82(1):8-24. [Medline: 11784274]
46. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39(11):897-906. [Medline: 2947907]
47. Grotle M, Brox JI, Vøllestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine (Phila Pa 1976) 2004 Nov 01;29(21):E492-E501. [Medline: 15507789]
48. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007 Jan;60(1):34-42. [doi: 10.1016/j.jclinepi.2006.03.012] [Medline: 17161752]
49. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989 Dec;10(4):407-415. [Medline: 2691207]
50. Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. theoretical considerations and an example application of change in health status. PLoS One 2014;9(12):e114468 [FREE Full text] [doi: 10.1371/journal.pone.0114468] [Medline: 25474472]
51. ROCMIC: Stata module to estimate minimally important change (MIC) thresholds for continuous clinical outcome measures using ROC curves database on the Internet. URL:https://ideas.repec.org/c/boc/bocode/s457052.html [accessed 2018-08-29] [WebCite Cache ID 722gkYSm4]
52. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979 Mar;86(2):420-428. [Medline: 18839484]
53. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006 Oct;59(10):1033-1039. [doi: 10.1016/j.jclinepi.2005.10.015] [Medline: 16980142]
54. Sitgreaves R. Review of Intraclass correlation and the analysis of variance by E. A. Haggard. J Am Stat Assoc 1960;55:384-385.
55. Fleiss JL. The Design and Analysis of Clinical Experiments. New York: Wiley; 1986.
56. Stratford PW, Riddle DL. A Roland Morris Disability Questionnaire target value to distinguish between functional and dysfunctional states in people with low back pain. Physiother Can 2016;68(1):29-35 [FREE Full text] [doi: 10.3138/ptc.2014-85] [Medline: 27504045]
57. Ostelo RWJG, de Vet HC. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 2005 Aug;19(4):593-607. [doi: 10.1016/j.berh.2005.03.003] [Medline: 15949778]
58. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002 Sep;55(9):900-908. [Medline: 12393078]
59. Tilford J, Roberson P, Fiser D. Using lfit and lroc to evaluate mortality prediction models. Stata Technical Bulletin 1995;5:77-81.
60. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1991.
61. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
62. Efron B, Tibshirani R. Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science (1) 1986:54-77.
63. Froud R. Improving Interpretation of Patient-Reported Outcomes in Low Back Pain Trials. London: Queen Mary University of London; 2010.
64. Gould W, Pitblado J. ACCUM: Stata module. 2001. URL:https://www.stata.com/support/faqs/statistics/bootstrapped-samples-guidelines/ [accessed 2018-01-12] [WebCite Cache ID 6wPzQpaXc]

65. Fawkes CA, Leach CMJ, Mathias S, Moore AP. Development of a data collection tool to profile osteopathic practice: use of a nominal group technique to enhance clinician involvement. Man Ther 2014 Apr;19(2):119-124. [doi: 10.1016/j.math.2013.08.006] [Medline: 24119310]
66. Fawkes CA, Leach CMJ, Mathias S, Moore AP. A profile of osteopathic care in private practices in the United Kingdom: a national pilot using standardised data collection. Man Ther 2014 Apr;19(2):125-130. [doi: 10.1016/j.math.2013.09.001] [Medline: 24139392]
67. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. Eur Spine J 2010 Sep;19(9):1484-1494 [FREE Full text] [doi: 10.1007/s00586-010-1353-6] [Medline: 20397032]
68. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998 Nov;78(11):1186-1196. [Medline: 9806623]
69. Beurskens AJ, de Vet HC, Köke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain: assessment of the quality of four disease-specific questionnaires. Spine (Phila Pa 1976) 1995 May 01;20(9):1017-1028. [Medline: 7631231]
70. Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index v2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. Spine (Phila Pa 1976) 2008 Oct 15;33(22):2450-2457. [doi: 10.1097/BRS.0b013e31818916fd] [Medline: 18824951]
71. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. Phys Ther 1994 Jun;74(6):528-533. [Medline: 8197239]
72. Mannion AF, Junge A, Grob D, Dvorak J, Fairbank JCT. Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. Eur Spine J 2006 Jan;15(1):66-73 [FREE Full text] [doi: 10.1007/s00586-004-0816-z] [Medline: 15856340]
73. Coelho RA, Siqueira FB, Ferreira PH, Ferreira ML. Responsiveness of the Brazilian-Portuguese version of the Oswestry Disability Index in subjects with low back pain. Eur Spine J 2008 Aug;17(8):1101-1106 [FREE Full text] [doi: 10.1007/s00586-008-0690-1] [Medline: 18512083]
74. Brouwer S, Kuijer W, Dijkstra PU, Göeken LNH, Groothoff JW, Geertzen JHB. Reliability and stability of the Roland Morris Disability Questionnaire: intra class correlation and limits of agreement. Disabil Rehabil 2004 Feb 04;26(3):162-165. [doi: 10.1080/09638280310001639713] [Medline: 14754627]
75. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Danish version of the Oswestry disability index for patients with low back pain. Part 2: Sensitivity, specificity and clinically significant improvement in two low back pain populations. Eur Spine J 2006 Nov;15(11):1717-1728. [doi: 10.1007/s00586-006-0128-6] [Medline: 16736202]
76. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain 2004 Aug;8(4):283-291. [doi: 10.1016/j.ejpain.2003.09.004] [Medline: 15207508]
77. Janwantanakul P, Sihawong R, Sitthipornvorakul E, Paksaichol A. A screening tool for non-specific low back pain with disability in office workers: a 1-year prospective cohort study. BMC Musculoskelet Disord 2015 Oct 14;16:298 [FREE Full text] [doi: 10.1186/s12891-015-0768-y] [Medline: 26467434]
78. Scrimshaw SV, Maher C. Responsiveness of visual analogue and McGill pain scale measures. J Manipulative Physiol Ther 2001 Oct;24(8):501-504. [doi: 10.1067/mmt.2001.118208] [Medline: 11677548]
79. Parker SL, Adogwa O, Paul AR, Anderson WN, Aaronson O, Cheng JS, et al. Utility of minimum clinically important difference in assessing pain, disability, and health state after transforaminal lumbar interbody fusion for degenerative lumbar spondylolisthesis. J Neurosurg Spine 2011 May;14(5):598-604. [doi: 10.3171/2010.12.SPINE10472] [Medline: 21332281]
80. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord 2006 Oct 25;7:82 [FREE Full text] [doi: 10.1186/1471-2474-7-82] [Medline: 17064410]
81. Kovacs FM, Abraira V, Royuela A, Corcoll J, Alegre L, Cano A, et al. Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. Spine (Phila Pa 1976) 2007 Dec 01;32(25):2915-2920. [doi: 10.1097/BRS.0b013e31815b75ae] [Medline: 18246018]
82. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 2. Phys Ther 1998 Nov;78(11):1197-1207. [Medline: 9806624]
83. Jordan K, Dunn KM, Lewis M, Croft P. A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. J Clin Epidemiol 2006 Jan;59(1):45-52. [doi: 10.1016/j.jclinepi.2005.03.018] [Medline: 16360560]
84. van der Roen N, Ostelo RWJG, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine (Phila Pa 1976) 2006 Mar 01;31(5):578-582. [doi: 10.1097/01.brs.0000201293.57439.47] [Medline: 16508555]
85. Rose G. The Strategy of Preventive Medicine. Oxford: Oxford University Press; 1992.
86. de Vet HC, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. J Rheumatol 2006 Feb;33(2):434-435. [Medline: 16465677]

87. Costa LOP, Maher CG, Latimer J. Self-report outcome measures for low back pain: searching for international cross-cultural adaptations. Spine (Phila Pa 1976) 2007 Apr 20;32(9):1028-1037. [doi: 10.1097/01.brs.0000261024.27926.0f] [Medline: 17450079]
88. Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Semmer NK, et al. Outcome assessment in low back pain: how low can you go? Eur Spine J 2005 Dec;14(10):1014-1026. [doi: 10.1007/s00586-005-0911-9] [Medline: 15937673]
89. Parker SL, Mendenhall SK, Shau DN, Adogwa O, Anderson WN, Devin CJ, et al. Minimum clinically important difference in pain, disability, and quality of life after neural decompression and fusion for same-level recurrent lumbar stenosis: understanding clinical versus statistical significance. J Neurosurg Spine 2012 May;16(5):471-478. [doi: 10.3171/2012.1.SPINE11842] [Medline: 22324801]
90. Bird M, Callisaya ML, Cannell J, Gibbons T, Smith ST, Ahuja KD. Accuracy, validity, and reliability of an electronic visual analog scale for pain on a touch screen tablet in healthy older adults: a clinical trial. Interact J Med Res 2016 Jan 14;5(1):e3 [FREE Full text] [doi: 10.2196/ijmr.4910] [Medline: 26769149]
91. Underwood MR, Parsons S, Eldridge SM, Spencer AE, Feder GS. Asking older people about fear of falling did not have a negative effect. J Clin Epidemiol 2006 Jun;59(6):629-634. [doi: 10.1016/j.jclinepi.2005.09.014] [Medline: 16713526]
92. Bijur PE, Silver W, Gallagher EJ. Reliability of the visual analog scale for measurement of acute pain. Acad Emerg Med 2001 Dec;8(12):1153-1157. [Medline: 11733293]
93. Gallagher EJ, Bijur PE, Latimer C, Silver W. Reliability and validity of a visual analog scale for acute abdominal pain in the ED. Am J Emerg Med 2002 Jul;20(4):287-290. [Medline: 12098173]
94. Bishop FL, Lewis G, Harris S, McKay N, Prentice P, Thiel H, et al. A within-subjects trial to test the equivalence of online and paper outcome measures: the Roland Morris disability questionnaire. BMC Musculoskelet Disord 2010 Jun 08;11:113 [FREE Full text] [doi: 10.1186/1471-2474-11-113] [Medline: 20529332]
95. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017 Apr 19;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
96. Mokkink LB, Prinsen CAC, Bouter LM, de Vet HC, Terwee CB. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther 2016 Jan 19;20(2):105-113 [FREE Full text] [doi: 10.1590/bjpt-rbf.2014.0143] [Medline: 26786084]
97. Froud R, Ellard D, Patel S, Eldridge S, Underwood M. Primary outcome measure use in back pain trials may need radical reassessment. BMC Musculoskelet Disord 2015 Apr 14;16:88 [FREE Full text] [doi: 10.1186/s12891-015-0534-1] [Medline: 25887581]
98. Fawkes C, Carnes D, Froud R. Introducing electronic PROM data collection into clinical practice. Learning the lessons from a pilot study. Physiotherapy 2017;103:e111.
99. Fawkes C. The Development, Evaluation, and Initial Implementation Of A National Programme For The Use and Collation Of Patient Reported Outcome Measures (Proms) In Osteopathic Back Pain Services In The UK. London: Queen Mary University of London; 2017.
100. Deeny SR, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. BMJ Qual Saf 2015 Aug;24(8):505-515 [FREE Full text] [doi: 10.1136/bmjqs-2015-004278] [Medline: 26065466]
101. Celi LA, Davidzon G, Johnson AE, Komorowski M, Marshall DC, Nair SS, et al. Bridging the health data divide. J Med Internet Res 2016 Dec 20;18(12):e325 [FREE Full text] [doi: 10.2196/jmir.6400] [Medline: 27998877]

Abbreviations

AUC: area under the curve
COSMIN: Consensus-Based Standards for the Selection of Health Measurement Instruments
eNRS: electronic numerical rating scale
ePRO: electronic patient-reported outcome measure
eRMDQ: electronic Roland Morris Disability Questionnaire
eVAS: electronic visual analog scale
ICC: intraclass correlation coefficient
MIC: minimally important change
MDC: minimal detectable change
MDC_{95}: minimal detectable change at the 95% level
NRS: numerical rating scale
PROM: patient-reported outcome measure
RMDQ: Roland Morris Disability Questionnaire
ROC: receiver operator characteristic
VAS: visual analog scale

Edited by G Eysenbach; submitted 15.01.18; peer-reviewed by M Alshehri, A Riis; comments to author 30.03.18; revised version received 18.05.18; accepted 18.06.18; published 24.10.18

Please cite as:
Froud R, Fawkes C, Foss J, Underwood M, Carnes D
Responsiveness, Reliability, and Minimally Important and Minimal Detectable Changes of 3 Electronic Patient-Reported Outcome Measures for Low Back Pain: Validation Study
J Med Internet Res 2018;20(10):e272
URL: http://www.jmir.org/2018/10/e272/
doi: 10.2196/jmir.9828
PMID: 30355556